DOCUMENT RESUME

ED 326 550 TM 015 809

AUTHOR Allen, David E.
TITLE A Simulation Investigation of Principal Component Regression.
PUB DATE 88
NOTE 17p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Computer Simulation; *Equations (Mathematics); *Factor Analysis; *Least Squares Statistics; *Predictor Variables; *Regression (Statistics)
IDENTIFIERS Eigenvalues; *Multicollinearity

ABSTRACT

Regression analysis is one of the more common analytic tools used by researchers. However, multicollinearity between the predictor variables can cause problems in using the results of regression analyses. Problems associated with multicollinearity include entanglement of relative influences of variables due to reduced precision of estimation, inappropriate dropping of variables from analysis, and oversensitivity of estimates of coefficients to particular sets of sample data. Alternative estimators have been developed that, when faced with severe multicollinearity, result in better estimates than do ordinary least squares (OLS) methods. However, little research has been conducted concerning the break-even point between OLS and other alternatives. The simulation presented in this article addresses this problem using regression that involves principal components. Data were generated for each of 72 experimental runs. Results indicate that the following three alternatives, differing on the criterion for selecting which components to use, appear to be reasonable: the Park criterion; the all but one criterion; and selection up to all but one based on correlation with the criterion variable. A 16-item list of references and 4 data tables are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED326550

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

DAVID E. ALLEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A SIMULATION INVESTIGATION OF PRINCIPAL COMPONENT REGRESSION

David E. Allen
Northern Michigan University

Address correspondence to:
David E. Allen
Management, Marketing and
Computer Information Systems
Northern Michigan University
Marquette, MI, 49855

TM015809

A SIMULATION INVESTIGATION OF PRINCIPAL COMPONENT REGRESSION

Regression analysis is one of the more common analytic tools used by researchers. However, multicollinearity between the predictor variables can cause problems in using the results of this analysis. Alternative estimators have been developed that, when faced with severe multicollinearity, result in better estimates than OLS. However, little investigation has been conducted concerning the breakeven point between OLS and the alternatives. This simulation addresses this problem, using as an alternative regression using principal components.

KEY WORDS: REGRESSION; MULTICOLLINEARITY; PRINCIPAL COMPONENTS; EIGENVALUES; ORDINARY LEAST SQUARES

A SIMULATION INVESTIGATION OF PRINCIPAL COMPONENT REGRESSION

Introduction

An investigation into the types of statistical techniques most often reported in five behavioral science journals indicated that regression/correlation analysis enjoyed a number one ranking in each journal, accounting for approximately 40% of the analyses conducted (Gaither and Glorfeld, 1985). The regression model is:

$$Y = X\beta + E \tag{1}$$

where: Y is an (n x 1) vector
X is an (n x p) matrix
β is the (p x 1) coefficient vector
E is the (n x 1) vector of unobserved errors

(with the usual assumptions of X being fixed, X and E independent, and E being independent and identically distributed error term with expected value of 0), the estimator of β is:

$$B = (X'X)^{-1}X'Y \tag{2}$$

The estimator (2) is unbiased with variances and covariances of the coefficients given by:

$$V(B) = \sigma^2(X'X)^{-1} \tag{3}$$

As a consequence of its popularity, OLS regression routines are found in all but the most primitive statistical software packages.

Often little direct attention is given to the problem of multicollinearity (a situation that occurs when the X-variables are interrelated). Multicollinearity presents problems (Johnston, 1970, p. 160):

1. The precision of estimation falls so that it becomes very difficult, if not impossible, to disentangle the relative influences of the X variables. Loss of precision has three aspects: specific estimates may have very large errors; these errors may be highly correlated, one with another; and the sampling variances of the coefficients will be very large.

2. Investigators are sometimes led to drop variables incorrectly from an analysis because their coefficients are not significantly different from zero, but the true situation may be not that the variable has no effect but simply that the set of sample data has not enabled us to pick it up.

3. Estimates of coefficients become very sensitive to particular sets of sample data, and the addition of a few more observations can sometimes produce dramatic shifts in some of the coefficients.

Multicollinearity is common. The real question is not whether we are faced with multicollinearity, but with the magnitude of it. While Mitchell (1985) reported that 67% of the studies he surveyed indicated that they tested for and dealt with the issue of multicollinearity, the commonly used procedures may not adequately address the problem.

Common ways of addressing problems of multicollinearity include ignoring the issue, dropping collinear variables from the analysis (usually the variables with the weakest theoretical support), using a stepwise regression technique, and controlling for effects with partial correlations (Green, 1978). While each of these methods is feasible in the sense that some of the problems induced by multicollinearity are reduced, each alternative can result in different outcomes. As stated by Intrilligator (1978), multicollinearity is not necessarily a serious problem if the purpose is to simply predict the dependent variable; however, if the goal of the study is to understand the influences of explanatory variables on the behavior of a dependent variable, the issue of multicollinearity cannot be ignored.

Various texts suggest that we either live with multicollinearity or discard offending variables (e.g., Green (1978) or Johnston (1970)). However, better solutions exist. A number of estimators that give biased results perform better in the sense of minimizing the mean squared error (variance plus squared bias) of the estimated coefficients (Gunst and Mason, 1977). One is the principal components estimator (PC), given by:

$$\hat{B}_{pc} = \sum_{j=1}^{r} c_j V_j / w_j \qquad r \leq p \qquad (4)$$
$$= V(Z'Z)^{-1}Z'Y$$

where:

w_j are the eigenvalues of $X'X$, $w_1 \geq w_2 \geq \ldots \geq w_p$

V_j are eigenvectors associated with w_j and V is (p x r)

$c_j = V_j'X'Y$

$Z_j = XV_j$ = principal components of $X'X^{-1}$ - Z is (n x r)

If r = p, then we have the principal component formulation of OLS. Note that the last (p-r) principal components, the ones with the smallest eigenvectors, are not included. This estimator is biased (Gunst and Mason, 1977):

$$E(\hat{B}_{pc}) = B - \sum_{j=r+1}^{p}(V_j'B)V_j \quad (5)$$

$$V(\hat{B}_{pc}) = \sigma^2\sum_{j=1}^{r}V_jV_j'/w_j \quad (6)$$

$$= V(\hat{B}_{ols}) - \sigma^2\sum_{j=r+1}^{p}V_jV_j'/w_j$$

$$MSE(\hat{B}_{pc}) = \sigma^2\sum_{j=1}^{r}1/w_j + \sum_{j=r+1}^{p}(V_j'B)^2 \quad (7)$$

The coefficients are biased towards zero. The variance of the estimated coefficients is certainly reduced. And with collinear data, the reduction in variance can overcome the bias, resulting in reduced mean squared error. These results are generally supported by simulation studies. While Dempster, Schatzoff and Wermuth (1977) report only minor and somewhat mixed results, Gunst and Mason (1977) criticize their use of stochastic selection of the number of components to be dropped, claiming the introduction of instability in the estimation. Gunst and Mason's results showed remarkable improvement over OLS in multicollinear situations.

Many other estimators have been identified as potential candidates for coping with multicollinearity (see, for example, Dempster, Schatzoff and Wermuth, 1977). Included would be various selection procedures for OLS (such as stepwise), and alternatives to OLS such as ridge regression, latent root regression, Stein regressors, etc. The OLS based selection procedures are generally inferior to the alternatives in terms of mean squared error; however, other methods may be superior to PC (this is an area needing further research). One of the more popular alternates, ridge regression, does have its problems; see Smith and Campbell for details. An important advantage of PC over the previously mentioned alternatives to OLS is that it is readily implemented using standard statistical packages such as SPSS or BMDP, as well as microcomputer software such as SYSTAT having both principal components and regression procedures:

Finally, the principal components are orthogonal to each other; thus it is easy to ascertain the marginal value of a specific component to the final regression since its impact does not depend on the other components in the regression.

There are a number of largely unresolved problems in the PC procedure. One concerns when to use a biased estimator instead of OLS. This is the purpose of the current investigation. While there are a number of ways to measure the amount of multicollinearity present, a particularly convenient measure is the condition number, usually presented as the ratio of the largest root to the smallest root:

$$\text{Cond } \# = w_1 / w_p \tag{8}$$

(The condition number computed from the eigenvalues of the correlation matrix R is a relatively unambiguous measure of multicollinearity. However, a condition number computed from the eigenvalues of X'X or Σ_X are confounded by the variances involved. It is entirely possible in the latter case to have a high condition number with iid variables.)

Another concerns the number of components to be retained. While in ordinary principal components analysis variable reduction is desired so as to enhance the interpretation of the relationships amongst the variables, the opposite appears to be true when using principal components solely to estimate regression models. In PC regression, components are dropped only if their associated eigenvalue is truly insignificant. Gunst and Mason (1977) showed significant improvement in mean squared error when only one component was dropped. Park (1981) states that the j-th principal component should be dropped if:

$$\sigma^2 - (\beta'\beta)w_j/p > 0 \tag{9}$$

Rearranging terms, we get:

$$w_j < p\sigma^2/\beta'\beta$$

Remember that the w_j were computed from $(X'X)^{-1}$; if computed from the variance covariance matrix Σ_X, the criterion is:

$$w_j < p\sigma^2/((N-1)\beta'\beta$$

Park recommends using the OLS estimates in computing (9).

Some interesting hypotheses can be developed from the last formulation of (9). We would expect the cutoff point to be larger (allowing easier entry of components into the regression) when we have more variables.

It would be smaller with more observations, and also would be smaller with a higher signal to noise ratio (note that $\beta'\beta/\sigma^2$ is the signal to noise ratio when $\Sigma_X = I$).

The spirit behind latent root regression would suggest using a stepwise procedure, entering only those components highly correlated with Y.

Somewhat related is the problem of which variables to use. So far, no stepwise procedure has been developed, and the researcher is left solely with a priori grounds for variable selection. A manual iterative backwards elimination procedure could be used to discard insignificant variables.

Experimental Design

A number of estimators were investigated. Since some have recommended using stepwise techniques, these OLS based estimators (forward addition, backward elimination and stepwise)were also included. The Principal component based estimators differed on the basis of criteria for the number of components. These included: eliminating the component with the smallest root; eliminating components with roots less than one (a commonly used procedure in Factor Analysis); using components satisfying Park's criterion; and using the p-1 components most related to Y. All of these four were also used with backward elimination of predictor variables. In summary, the twelve estimators used were:

OLS-DRT - Ordinary least squares
OLS-BAK - OLS based backward elimination
OLS-FWD - OLS based forward addition
OLS-STP - OLS based stepwise regression
PCR-F-1 - (p-1) principal components
PCR-F-E - components with roots ≥ 1
PCR-F-P - components satisfying Park's criterion
PCR-F-S - stepwise addition of up to (p-1) components
PCR-B-1 - (p-1) principal components, backward elimination
PCR-B-E - components with roots ≥ 1, backward elimination
PCR-B-P - Park's criterion, backward elimination
PCR-B-S - stepwise, up to (p-1) components, backward elimination

The following experimental predictor variables were used:

IDR: (= 0 if designed $R^2 = .25$, = 1 if designed $R^2 = .75$) The two values of R^2 were selected to represent typically low values in social science research and typically high values.

IDN: (= 0 for sample size N = 40, = 1 for N = 100) For the number of predictor variables used in the regressions, N = 40 is on the small side, N = 100 is quite a bit more satisfactory.

M: (= 5, 6, 7) Number of predictor variables

COND: Correlation-based condition number. The six designed condition numbers (condition numbers of the input correlation matrices to the data generator) ranged from 1 (iid variables) to 1,200 (significant collinearity)

The first three were run on a full-factorial design, with COND ranging relatively uniformly over its range. Thus the resulting predictor variable matrix had meaningless collinearity.

Programs to generate the data, do the regressions, and other miscellaneous data handling chores were written in MS-FORTRAN.

Data was generated for each of the 72 experimental runs. Each run consisted of three generated predictor variable sets, with each predictor variable set having two replications of Y generated. Generation was done using a well-tested random number generator, transformation of variates to Normal using the Polar method, and using Cholesky decomposition of the input correlation matrix to generate the predictor variables. Output variation of Y-errors was held (in probability) to $\sigma^2 = 1$. In all cases, the regression coefficients equaled one.

Each run was estimated using the twelve estimators. Output for further analysis included:

BIAS: On each run, the average bias $(b_i - \beta_i)$ was computed.

RMSE: On each run, the average root mean squared error of coefficient estimation was computed

RE: Root of relative efficiency, using OLS as base. $RE = RMSE_i / RMSE_{ols}$

DIFF: $RMSE_{ols} - RMSE_i$

Thus much of the analysis is conducted comparative to OLS. In all, 432 data sets were analyzed by the twelve estimators.

The resulting output was further analyzed using SYSTAT as well as some miscellaneous MS-FORTRAN programs.

Analysis of Experiments

Table 1 shows sample means of the 432 runs. The mean RMSE for OLS was 0.839. On average, none of the eleven alternatives is better than OLS (average RE's greater than one, average DIFF negative). All of the eleven are biased towards zero, as expected. However, a run by run comparison was made between all twelve of the estimators; no one estimator dominated any other.

Table 2 compares RMSE's of the estimators, categorized by condition number groupings. The first group is characterized by extremely modest to no multicollinearity problems, the second group is modest to moderate multicollinearity, while the third group is moderately severe to severe. In evaluating the results, care should be taken. The standard errors of the means are not negligible.

However, it is quite clear that the "eigenvalue ≥ 1" criteria is not very good. What might be suitable for Factor Analysis is not suitable for regression.

In situations of lower multicollinearity, the two Park criterion estimators are essentially complex calculating clones of the ordinary least squares estimators OLS-DRT and OLS-BAK. The criteria sets severe standards before allowing a reduction in the number of components. In general, it seems to do a decent job.

The stepwise selection of components seems to do a better job than simply eliminating the component with the lowest root.

With regards to stepwise regressions, note that this study was biased against the selection of variables question; all variables had the same true coefficient value of one. However, the assertion that stepwise regression techniques are of value in reducing collinearity problems is called into question by their relatively poor performance in all condition number categories, including the last.

The results of this table seems to indicate that consideration should be given to OLS-DRT, PCR-F-1, PCR-F-P and PCR-F-S in estimating models that the researcher is confident of the specification (as mentioned above,

junk variables were not included in this research, and in the poorly specified case no recommendations can be given). Of these four, worst case scenarios suggest that estimation would be only about 20% more variable by selecting the worst instead of the best.

Tables 3 and 4 show results of regressions of either the difference between OLS-DRT and the estimator in question (table 3) or the natural log of root relative efficiency (table 4) versus noise, sample size, number of predictor variables and condition number. Since the Park criteria resulted in so many duplications of OLS-DRT, it was also estimated conditionally on differing from OLS-DRT. All 24 equations are statistically significant. But meaningful? The R''s are low. The attempt to find out exactly when to use a particular estimator was unfortunately swamped in estimation variability.

In 16 of the 24 regressions, noise (IDR) was significant. When not significant, the R^2 was quite low. Further, in all significant cases, the sign was in the direction to support the assertion that as the signal to noise ratio increases, we should be more hesitant about abandoning OLS.

In all regressions, sample size was significant. Further, in all cases, sign was in the direction to support the assertion that smaller sample sizes favor the PCR estimators, larger sample sizes favor inclusion of more components.

In only 13 regressions was the number of predictor variables significant. However, in all cases (significant or not) the sign was in the direction to assert that fewer variables favor OLS-DRT, more favor the alternatives.

In all but two cases as the log of the condition increased, the estimator in question became a better competitor to OLS-DRT. Further, these two cases were for the Park criterion estimated unconditionally; the large numbers of LNRE and DIFF = 0 swamped the other cases.

It would be foolhardy, in light of this data, to compute actual cutoff points and then use them in a rigid fashion. As an illustration,

however, consider PCR-F-1:

LNRE = 1.301 +0.171 IDR +0.557 IDN -0.151 M -0.110 LNC

DIFF = -1.621 -0.324 IDR -0.765 IDN +0.149 M +0.198 LNC

with high noise (IDR = 0), low sample size (IDN = 0) and 6 variables. In either case, the cutoff value of LNC is associated with LNRE = 0 or DIFF = 0. Solving, $LNC_{lnre} = 3.59$ and $LNC_{diff} = 3.67$; this yields cutoff condition numbers of 36 and 39 respectively.

Conclusions

This simulation study attempts to shed more information on when to use principal components based alternatives to ordinary least squares regression. It shows that three alternatives, differing on the criterion for selecting which components to use, appear to be reasonable: the Park criterion, the all but one criterion, and selecting up to all but one based on correlation to the criterion variable.

The study is limited in that only a particular coefficient pattern was used (all 1's). This biased the results against stepwise methods, which may be helpful in eliminating insignificant variables. Also, other methods (other than the condition number) may be more helpful in measuring multicollinearity. While the condition number may have flaws, it certainly is convenient. The same may be said for other alternatives to OLS other than the PC based estimators.

REFERENCES

Campbell J., Daft R. and Hulin, C. What to Study: Generating and Developing Research Questions. Beverly Hills: Sage, 1982.

Dagpunar, John, (1988) Principles of Random Variate Generation, New York: Oxford University Press.

Dempster, A. P., Schatzoff, M. and Wermuth, N. "A simulation study of alternatives to ordinary least squares," Journal of the American Statistical Association, 1977, 72, 77-90.

Draper, N. R. and Smith, H. Applied Regression Analysis. New York: John Wiley, 1982, 1966.

Gaither, N. and Glorfeld, L. "An evaluation of the use of tests of significance in organizational behavior research", Academy of Management Review, 1985, 10, 787-793.

Green, P. E. Analyzing Multivariate Data, Hinsdale, Ill.: Dryden Press, 1978.

Gunst, R. F. and Mason, R. L. "Biased estimation in regression: an evaluation using mean squared error," Journal of the American Statistical Association, 1977, 72, 616-28.

Gunst, R. F. and Mason, R. L. "Some considerations on the evaluation of alternative prediction equations," Technometrics, 1979, 21, 55-64.

Intrilligator, M. Econometric Models, Techniques and Applications. Englewood Cliffs, N. J.: Prentice-Hall, 1978.

Johnston, J. Econometric Methods, New York: McGraw-Hill, 1970.

Kerlinger, F. N. Foundations of Behavioral Research, New York: Holt, Rinehart and Winston, 1973.

Mansfield, E. R., Webster, J. T. and Gunst, R. F. "An analytic variable selection technique for principal component regression," Applied Statistics, 1977, 26, 34-40.

Mitchell, T. "An evaluation of the validity of correlational research conducted in organizations", Academy of Management Review, 1985, 10, 192-205.

Neter, J., Wasserman, W. and Kutner, M. Applied Linear Statistical Models, Homewood, Ill.: Richard D. Irwin, 1985.

Park, S. H. "Optimal restrictions on Regression Parameters," Technometrics, 1981, 23, 289-95.

Smith, G. and Campbell, F. "A critique of some ridge regression methods," Journal of the American Statistical Society, 1980, 75, 74-81.

Table 1: Sample Means

	RMSE	RE	LNRE	DIFF	BIAS
OLS-FWD	1.175	1.762	0.319	-0.336	-0.596
OLS-BAK	1.070	1.540	0.263	-0.231	-0.396
OLS-STP	1.183	1.771	0.325	-0.344	-0.609
PCR-F-1	1.195	1.695	0.250	-0.355	-0.589
PCR-F-E	1.863	2.871	0.478	-1.024	-1.556
PCR-F-P	0.881	1.177	0.077	-0.042	-0.220
PCR-F-S	0.872	1.261	-0.029	-0.033	-0.357
PCR-B-1	1.384	2.079	0.397	-0.545	-0.901
PCR-B-E	1.815	2.793	0.472	-0.976	-1.516
PCR-B-P	1.053	1.506	0.244	-0.213	-0.424
PCR-B-S	0.946	1.367	0.033	-0.107	-0.428

RMSE = Root Mean Squared Error, averaged
RE = RMSEest / RMSEols
LNRE = natural log (RE)
DIFF = RMSEols - RMSEest
BIAS = Est coef - True Coef, averaged over the m variables

Table 2: Mean RMSE by Condition Number

COND:	< 50	50 - 300		> 300	
N:	144		116		172
OLS-DRT	0.934	OLS-DRT	0.846	PCR-F-S	0.638
	(.029)		(.037)		(.028)
PCR-F-P	0.934	PCR-F-P	0.908	PCR-F-1	0.677
	(.029)		(.039)		(.023)
PCR-F-S	1.117	PCR-F-S	0.913	PCR-B-S	0.680
	(.032)		(.050)		(.032)
PCR-B-S	1.228	PCR-B-S	0.991	OLS-DRT	0.755
	(.040)		(.055)		(.032)
PCR-B-P	1.248	PCR-F-1	1.038	OLS-FWD	0.813
	(.040)		(.048)		(.030)
OLS-BAK	1.253	PCR-B-P	1.132	PCR-F-P	0.818
	(.040)		(.050)		(.029)
PCR-F-1	1.939	OLS-BAK	1.136	OLS-STP	0.819
	(.106)		(.050)		(.031)
OLS-FWD	1.491	PCR-B-1	1.300	PCR-B-P	0.835
	(.061)		(.078)		(.030)
OLS-STP	1.494	OLS-FWD	1.320	PCR-B-1	0.870
	(.061)		(.067)		(.046)
PCR-B-1	2.066	OLS-STP	1.339	OLS-BAK	0.872
	(.096)		(.067)		(.031)
PCR-B-E	2.686	PCR-B-E	1.802	PCR-B-E	1.095
	(.129)		(.141)		(.072)
PCR-F-E	2.771	PCR-F-E	1.820	PCR-F-E	1.131
	(.135)		(.143)		(.073)

(std. error of mean)

Table 3. REGRESSIONS OF LNRE VS IDR, IDN, M, LNC

	R^2	σ	F	POS	CONST	IDR	IDN	M	LNC
OLS-FWD	.098	0.652	11.574	0.000	0.645	0.114	0.355	-0.062	-0.048
					(2.592)	(1.819)	(5.658)	(1.605)	(2.837)
OLS-BAK	.034	0.554	3.749	0.005	0.574	0.111	0.131	-0.064	-0.011
					(2.716)	(2.076)	(2.452)	(1.959)	(0.850)
OLS-STP	.096	0.652	11.382	0.000	0.595	0.121	0.354	-0.054	-0.040
					(2.392)	(1.929)	(5.642)	(1.397)	(2.759)
					(5.911)	(3.071)	(10.046)	(4.447)	(8.548)
PCR-F-E	.288	1.040	43.262	0.000	1.937	0.379	1.004	-0.245	-0.148
					(4.881)	(3.785)	(10.034)	(3.987)	(6.388)
PCR-F-P	.090	0.341	10.534	0.000	0.347	-0.007	0.143	-0.074	0.023
					(2.667)	(0.221)	(4.351)	(3.668)	(2.988)
PCR-F-S	.065	0.765	7.469	0.000	0.521	-0.000	0.271	-0.064	-0.068
					(1.783)	(0.003)	(3.674)	(1.414)	(3.828)
PCR-B-1	.266	0.701	38.780	0.000	1.634	0.263	0.597	-0.206	-0.094
					(6.106)	(3.896)	(8.847)	(4.978)	(5.990)
PCR-B-E	.294	1.000	44.494	0.000	1.862	0.379	0.976	-0.233	-0.145
					(4.881)	(3.934)	(10.142)	(3.948)	(6.520)
PCR-B-P	.035	0.551	3.887	0.004	0.585	0.114	0.119	-0.063	-0.017
					(2.783)	(2.152)	(2.236)	(1.943)	(1.394)
PCR-B-S	.070	0.804	8.040	0.000	0.628	0.064	0.265	-0.069	-0.075
					(2.049)	(0.821)	(3.423)	(1.450)	(4.193)
Conditional: N = 136									
PCR-F-P	.172	0.556	6.808	0.000	1.805	0.045	0.394	-0.172	-0.131
					(3.619)	(0.444)	(3.918)	(2.879)	(2.195)

(abs t-value)
N = 432
LNRE = natural log (RE)
RE = RMSEest / RMSEols
RMSE = Root Mean Squared Error, averaged over the m variables
IDR = 0 if R2 = .25; = 1 if R2 = .75
IDN = 0 if N = 40; = 1 if N = 100
M = number of predictor variables = 5, 6, 7
LNC = natural log (COND)
COND = Condition number = EV1 / EVm
Conditional PCR-F-P: When less than m components entered

Table 4: REGRESSIONS OF DIFF VS IDR, IDN, M, LNC

	R^2	σ	F	POS	CONST	IDR	IDN	M	LNC
OLS-FWD	.157	0.659	19.838	0.000	-0.705	-0.153	-0.420	0.051	0.076
					(2.807)	(2.417)	(6.629)	(1.310)	(5.170)
OLS-BAK	.070	0.455	8.081	0.000	-0.567	-0.127	-0.113	0.048	0.037
					(3.268)	(2.895)	(2.588)	(1.780)	(3.634)
OLS-STP	.156	0.659	19.681	0.000	-0.641	-0.162	-0.421	0.040	0.075
					(2.548)	(2.543)	(6.638)	(1.041)	(5.084)
PCR-F-1	.392	0.768	68.744	0.000	-1.621	-0.324	-0.765	0.149	0.198
					(5.536)	(4.376)	(10.35)	(3.298)	(11.57)
PCR-F-E	.556	1.048	133.61	0.000	-2.633	-0.757	-1.810	0.278	0.265
					(6.586)	(7.498)	(17.95)	(4.506)	(11.33)
PCR-F-P	.066	0.242	7.574	0.000	-0.248	0.001	-0.092	0.049	-0.009
					(2.689)	(0.032)	(3.961)	(3.416)	(1.603)
PCR-F-S	.085	0.470	9.918	0.000	-0.367	-0.024	-0.175	0.034	0.050
					(2.050)	(0.530)	(3.880)	(1.223)	(4.777)
PCR-B-1	.392	0.813	68.787	0.000	-2.214	-0.406	-0.870	0.249	0.176
					(7.138)	(5.183)	(11.11)	(5.190)	(9.730)
PCR-B-E	.549	1.025	130.13	0.000	-2.502	-0.751	-1.744	0.267	0.254
					(6.399)	(7.600)	(17.68)	(4.415)	(11.11)
PCR-B-P	.077	0.453	8.925	0.000	-0.576	-0.135	-0.093	0.047	0.042
					(3.336)	(3.089)	(2.145)	(1.767)	(4.164)
PCR-B-S	.112	0.516	13.509	0.000	-0.536	-0.078	-0.191	0.042	0.067
					(2.723)	(1.560)	(3.850)	(1.384)	(5.847)
Conditional (N=136)									
PCR-F-P	.213	0.389	8.863	0.000	-1.598	-0.047	-0.266	0.121	0.149
					(4.582)	(0.663)	(3.775)	(2.900)	(3.575)

(abs t-value)
N = 432
DIFF = RMSest - RMSols
RMSE = Root Mean Squared Error, averaged over the m variables
IDR = 0 if R2 = .25; = 1 if R2 = .75
IDN = 0 if N = 40; = 1 if N = 100
M = number of predictor variables = 5, 6, 7
LNC = natural log (COND)
COND = Condition number = EV1 / EVm
Conditional PCR-F-P: When less than m components entered